

# Non-Thermal Transitions in $n$ -th Order Moral Decisions

Roberto C. Alamino

*Non-linearity and Complexity Research Group,*

*Aston University, Birmingham B4 7ET, UK*

## Abstract

This work introduces a model in which agents of a network act upon one another according to three different kinds of moral decisions. These decisions are based on an increasing level of sophistication in the empathy capacity of the agent, a hierarchy which we name *Piaget's Ladder*. The decision strategy of the agents is non-rational, in the sense that it does not minimize model's Hamiltonian, and the model presents quenched disorder given by the distribution of its defining parameters. We obtain an analytical solution for this model in the thermodynamic limit and also a leading order correction for finite sized systems. Using these results, we show that typical realizations develop a rich phase structure with discontinuous non-thermal transitions.

## I. INTRODUCTION

Human societies are inherently complex. So much that recent research indicates that even highly specialized qualitative knowledge from social sciences does not improve prediction performance on average [1]. On the other hand, statistical physics models have been successful in reproducing and predicting observed emergent behavior in real social datasets [2–5], giving birth to a whole new branch of physics, still in its infancy, which has been named *Sociophysics* [6].

Social scientists and psychologists often see these results with suspicion, arguing that each human individual is different. It is however clear that human behavior presents identifiable patterns without which their very disciplines would not exist. The difficulty comes from the numerous sources of disorder which can affect individual behavior and their decisions concerning social interactions [7, 8] as, for instance, the influence of mass media [9] or the availability of natural resources [10]. The contribution of statistical physics has been in understanding how emergent behaviors depend on the disorder and which are the relevant parameters driving transitions between the different phases [11].

One of the most studied problems in sociology concerns the emergence of moral behavior and its consequences to the general well-being and survival ability of a certain population [12]. There is evidence that moral behavior is the result of multi-level selection [13, 14], but the exact mechanism is still not understood. Although the picture is not yet complete, some of its parts are gradually becoming clearer. For instance, the essential role of intuitive (emotional) judgments in the process of moral formation and decision making is now well accepted and has been recently used in the formulation of the framework known as Moral Foundations Theory (MFT) [15].

The essential role of emotions is, in fact, nothing but expected as motivation from emotional satisfaction, independently of cultural differences, is a factor that has been identified a long time ago as being decisive in the survival of an organism or a group even when their physical needs are fulfilled [16].

In this work we use a simplified statistical mechanics model to study the influence of different moral decisions in the overall emotional satisfaction of a human population. We will not be interested in the problem of emergence of particular moral beliefs and, therefore, we will assume that the concepts of what is morally acceptable is agreed *a priori* by the

whole population. It is obvious that these concepts will vary and even exchange places in different societies, but this will not affect our analysis.

The model analyzed here is infinite dimensional in the sense that each person from the group interacts with every other person by taking a one-time binary decision: to act *helpfully* or *harmfully* relative to the group’s agreed moral rules. The person’s decision is assumed to be a *conscious* one, which is why we can actually call it a moral decision. We then classify the moral decisions according to the level of empathy embodied in them.

Although MFT suggests that humans classify moral beliefs using at least five dimensions instead of only one binary dimension (harmful/helpful), this simplification is enough for the purposes of this work. A more sophisticated model can be obtained by considering the agents as neural networks classifying moral multidimensional binary vectors as in [4].

The spectrum of moral decisions that can be taken by humans can be extremely complicated and analyzing it would be out of the scope of this paper. Here we focus on what we call  $n$ -th order moral decisions, where the order of the decision identifies the increasing level of empathy necessary to take it according to the work of the well-known psychologist Jean Piaget [17]. Piaget observed that there are distinct cognitive stages in the development of the child intellect before it reaches the adult stage. This evolution occurs in three steps with increasing levels of empathic capacity. It is this sequence of three steps that we call *Piaget’s Ladder* and which defines the order of a moral decision.

Piaget has proposed, based on empirical observations, that every children first develops a sense of self in which it is capable of understand its own feelings, but is unable to recognize the feelings of others, acting only selfishly. Accordingly, we call this step the Selfish Step and decisions taken selfishly are considered of 0th order. As its development proceeds, the child becomes capable of recognizing that others also have feelings, but it is still unable to see things from others’ perspective. We call it the Parental Step (1st order) as it is not uncommon to parents to project their desires in the way the act towards their children. Finally, the cognitive abilities of the child reach a stage in which it can finally understand that others have different needs. This is the Empathic Step (2nd order). The details of Piaget’s original theory of cognitive development have been several times revised to account for further experimental evidence [18], but the details will not be as important for our purposes, only its key idea of incremental empathy.

The agents of our network will act on one another using moral decisions which are com-

binations of these three steps. These decisions, although conscious, are not rational in the sense that agents do not devise strategies to maximize their well-being (to be defined rigorously later) and simply follow a limited set of pre-determined rules. It is well known that in a varied number of scenarios, this is usually the rule rather than the exception [19]. The average well-being of the group defines then the phases of the statistical physics model and we are then interested in characterizing how they change as the set of disorder parameters of the model is varied.

A detailed explanation of how the model is constructed will be given in Sec. II together with its analytical solution. The phase structure of this solution is obtained and analyzed in Sec. III. Finally, we present our conclusions and further discussion in Sec. IV.

## II. THE MODEL

We consider a population of  $N$  agents represented by the nodes of a fully connected social network, as the one shown in fig.1 (left). Links between nodes indicate direct interactions between agents. As the aim of this work is primarily to identify the emergence of collective behavior in general, we solve only the fully connected case. The model can be straightforwardly generalized to any topology, with an obvious increased difficulty for obtaining exact solutions and we therefore will leave their analysis to forthcoming papers.

In the thermodynamic limit of  $N \rightarrow \infty$  (in which the fully connected case flows to the mean-field solution) the model develops a very rich phase structure with different kinds of non-thermal transitions. These transitions are induced by varying the intensity of the disorder, understood to be the parameters of the different probability distributions of the model which we will describe in the following.

Each static configuration of the network is defined by one single (frozen) realization of three variables, namely  $J$ ,  $\mathbf{u}$  and  $\mathbf{v}$ . The variable  $J$  is a matrix containing the moral decisions taken by the agents towards others. A certain realization of  $J$  is obtained by each agent on a generic site  $i$  acting on another agent at site  $j$  by taking a one-time binary decision: it either acts *harmfully* or *helpfully* towards  $j$ , where these notions are judged according to *a priori* moral concepts agreed by the entire network. Notice that neither choice is directly associated to objective physical or psychological harm, but is based on the subjective acceptance of this actions in the agent's network (group, community, population, culture etc).

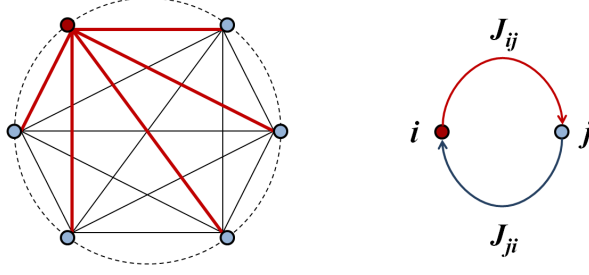


FIG. 1. Fully connected social network (left) with non-symmetrical interactions between agents (right).

The action of  $i$  on  $j$  is then represented by the matrix element  $J_{ij} \in \{0, \pm 1\}$ , where  $-1$  indicates a *harmful* action and  $+1$  a *helpful* one. The zero value corresponds to no action at all. In the present model, we will restrict this value only to the diagonal entries of the matrix  $J$ , i.e., those of the form  $J_{ii}$  which represent self-interactions. The matrix  $J$  is *not symmetric* in general as the order of indexes *is* important, the first representing the agent that is originating the action and the second the one being targeted by it as depicted in fig.1 (right).

To each individual or node of the network, we associate a *personality vector*  $\pi_i = (u_i, w_i)$  with binary components  $u_i, w_i \in \{\pm 1\}$ . These two components represent “emotional desires” of the individual  $i$ . The first component,  $u_i$ , indicates how the agent  $i$  would “desire” to act towards another agent. When  $u_i = -1$ , this indicates that agent  $i$  prefers to do harmful actions, while a value  $+1$  would indicate a preference for helpful ones. The assumed *emotional* nature of these preferences is intended to mean that fulfilling them leads to a subconscious satisfaction of the agent, which one can roughly associate to human feelings. A sadistic personality, for instance, would be represented by  $u_i = -1$  as the agent feels pleasure by harming others, while  $u_i = +1$  would represent an altruistic agent who enjoys helping its peers.

The second component of the personality vector represents how the agent would like to be treated by others. A  $w_i = +1$  is what one would expect from someone who would appreciate the help of others, a behavior probably seen as “normal”, while  $w_i = -1$  corresponds to an agent that prefers to be mistreated, as would be the case of a “masochist” agent, for instance.

There are clearly several simplifications in the above personality attributions compared

to real situations. First, we consider that both  $u_i$  and  $w_i$  depend only on the agent  $i$ . They are the same independently of which agent is the target of the action. A more sophisticated version of this model would consider that this preference might vary according to the target and the agent's feelings towards it. For instance,  $i$  might be emotionally led to act differently whether it is interacting with its own mother or with a well-known serial killer.

Another simplification, which is less important for the situation we are going to study, concerns the fact that the personality vector might change with time. Individual emotional responses are not only genetically defined, but are affected by interaction with the environment. The scenario studied in this work does not deal with dynamics and, therefore, this simplification is not relevant in the present case. It is however improbable that a significant change in behavior that can affect the whole population occurs in a short time scale (although it is not impossible to happen) and, therefore, we believe that we can consider stable personalities for a certain macroscopic period of time as a first approximation.

The use of binary vectors to characterize the agents' personalities represents also a strong simplification. Human behavior has a wide spectrum of variability which would be better modeled by continuous rather than binary variables. The intention of the present model, however, is to take a first step towards this modeling and try to identify some general basic principles and emergent behaviors. It is well-known that simplified models are used even in social sciences and, in particular, in psychology, with the best known example being the Myers-Brigg Type Indicator [20] which classifies human behavior by considering only 14 types and is largely used by institutions to actually select prospective employees.

Given the two *personality components* of  $\pi_i$ , whether or not agent  $i$  feels fulfilled by the interactions within its network will be represented by its *satisfaction*

$$\sigma_i = \text{sgn}[\Omega(\pi, J, \gamma)], \quad (1)$$

with

$$\Omega(\pi, J, \gamma) = \frac{1}{N}[\gamma U_i + (1 - \gamma)W_i], \quad (2)$$

and

$$U_i = u_i \sum_{j \neq i} J_{ij}, \quad W_i = w_i \sum_{j \neq i} J_{ji}. \quad (3)$$

This definition allows three values for the agent's satisfaction. When  $\sigma_i = +1$ , we say that the agent feels satisfied, when  $\sigma_i = -1$  it feels dissatisfied and when  $\sigma_i = 0$  the agent is neither.

The scaling  $1/N$  in  $\Omega$  is used to make it an intensive quantity in the number of agents, keeping it finite when  $N \rightarrow \infty$ . This is the limit in which we are interested as it is only when the number of agents is *exactly* infinite that we can unambiguously identify the model's phase transitions.

The extensive quantity  $U_i$  is the sum of all contributions to the fulfillment of the agent's desire on how to treat other agents, while the also extensive  $W_i$  represents the same concerning how the agent feels treated by the others. The parameter  $\gamma$  is taken from the interval  $[0, 1]$  for convenience and represents the relative *emotional* importance given by the individual to each of these terms. This parameter will be kept fixed and will be the same for the whole network.

We recall that in the static version analyzed here, each realization of the whole experiment will be considered as consisting of a fixed configuration of  $J$  and  $\pi$  which however varies from one realization to another. Random disorder enters the model through the distributions of possible values of these two variables in each realization of the “experiment” which might involve a totally different set of individuals and interactions each time.

Notice that the satisfaction contains only “emotional” contributions. Human beings have mechanisms to suppress emotional responses under rational considerations. Therefore, a term based on each agent's level of rationality can be devised which would compete with its emotional satisfaction. However, such a consideration is out of the scope of the present paper and should be addressed in more refined versions of the model.

For the present calculation, we assume that the personality parameters are i.i.d. with distribution

$$\mathcal{P}(u) = (1 - s)\delta(u, 1) + s\delta(u, -1), \quad (4)$$

$$\mathcal{P}(w) = (1 - m)\delta(w, 1) + m\delta(w, -1), \quad (5)$$

and  $s, m \in [0, 1]$  ( $s$  and  $m$  the same for all agents).

A natural Hamiltonian for the whole network would be the negative of its net satisfaction  $H = -\sum_i \sigma_i$  with the distribution of personalities being the quenched disorder and the moral decisions  $J_{ij}$  the dynamical variables (analogous to the spins in a magnetic model). This would lead to a thermodynamic distribution for  $J$  at inverse temperature  $\beta$  given by  $\mathcal{P}(J|\pi) \propto e^{-\beta H}$ . Such an equilibrium distribution corresponds to a dynamics in which agents are driven to act towards the minimization of the network's energy, a behavior that can be

$J_{ij}$	Order	Moral Behavior
$u_i$	0	Selfish
$w_i$	1	Parental
$w_j$	2	Empathetic

TABLE I.  $n$ -th order morals

classified as *rational*. As discussed in the Introduction, there are several reasons why agents might not act rationally. For instance, they might lack either the will or the resources to follow the strategy based on minimizing the energy. This is actually the case for our model and, therefore, the distribution of decisions will not be the equilibrium one for this natural choice of Hamiltonian.

In the particular case in which every agent has the same  $\gamma$ , the strategies that minimize the Hamiltonian are very simple and easy to see. If  $\gamma > 1/2$ , then more importance is given for  $U_i$  than for  $W_i$  and it benefits a totally selfish strategy. The opposite is true if  $\gamma < 1/2$ . These two strategies, however might not be optimal in general outside these parameter ranges. In fact, we will show that there are phase transitions in the average satisfaction of the networking (to be defined in the following) when using these strategies outside their optimality zone.

The specific distribution of the agent's actions  $J_{ij}$  we are going to use is, in fact, a key feature of this work. In addition to the thermal equilibrium distribution and the two simplified strategies considered above, there is clearly an infinite number of ways for an agent to choose how to act towards another one. For instance, all agents could simply act in the same way by choosing  $J_{ij} = +1$  independently of  $\pi_i$  or  $\pi_j$ . In order to be able to relate our work to empirical facts, we focus on what we described in the introduction as  $n$ -th order moral decisions based on Piaget's Ladder, which result in the values for the matrix  $J$  given in table I.

Here we consider the distribution of moral decisions defining the action of agent  $i$  towards agent  $j$  to be given by a convex combination of the  $n$ -order moral decisions as

$$\mathcal{P}(J_{ij}|\pi_i, \pi_j) = p_0\delta(J_{ij}, u_i) + p_1\delta(J_{ij}, w_i) + p_2\delta(J_{ij}, w_j), \quad (6)$$

with

$$\sum_i p_i = 1, \quad (7)$$



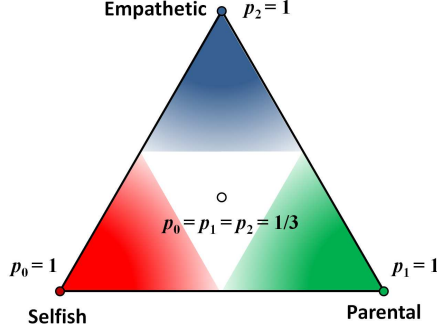


FIG. 2. Graphical representation of moral strategies as points of a triangle.

which can be represented as points in an equilateral triangle as in fig. 2.

The probabilities in the above equation can be seen either as the proportion of individuals choosing one of the three levels of moral decision or as the probability of one individual taking each of them at a certain time. Because the 2nd order decision requires knowledge of the personality vector of others, we will assume that agents possess full noiseless information about the  $w$  component of all other agents. A more realistic scenario would be when only partial information is available, adding an extra source of disorder.

The *average satisfaction* of the whole network can then be measured by averaging the individual satisfactions over the decision strategy and over the disorder in the personality vectors as

$$S = \langle \sigma_k \rangle_{\mathbf{u}, \mathbf{w}, J}, \quad (8)$$

in which the index  $k$  inside the average is only for convenience as it disappears due to the averaging over all variables  $u_i$ ,  $w_i$  and  $J_{ij}$  represented by the vectors  $\mathbf{u}$  and  $\mathbf{w}$  and the matrix  $J$ . This is equivalent to averaging over frozen realizations of the system.

We will take  $S$  as the order parameter defining the phases of the model. Ordered phases correspond to situations where  $S = +1$ , with all agents but for a set of measure zero are satisfied, and  $S = -1$ , when all agents are dissatisfied. Intermediate values of  $S$  correspond to disordered phases. In the next section, we analyze the different behavior of the system as the parameters of the model are varied.

### III. PHASE STRUCTURE

In the limit  $N \rightarrow \infty$  of an infinitely sized network, one can obtain the exact value of the average satisfaction for general values of the strategy's probabilities as (see appendix A for details)

$$S = \langle \text{sgn } \mu \rangle_{u,w}, \quad (9)$$

where

$$\mu = p_0[\gamma + (1 - \gamma)\bar{u}w] + p_1[\gamma uw + (1 - \gamma)\bar{w}w] + p_2[\gamma \bar{w}u + (1 - \gamma)], \quad (10)$$

and

$$\bar{u} = \langle u \rangle = (1 - 2s), \quad \bar{w} = \langle w \rangle = (1 - 2m). \quad (11)$$

In the following, we will analyze the four special cases represented by the points in fig. 2.

#### A. 0th Order Moral

When  $p_0 = 1$  then  $p_1 = p_2 = 0$  and we are assuming that all agents are acting selfishly, or taking 0th order moral decisions. Agents act by ignoring other agents' desires, maximally satisfying their own  $U_i$  in every decision. The expression for the average satisfaction simplifies to

$$S = (1 - m) \text{sgn}[\gamma + (1 - \gamma)(1 - 2s)] + m \text{sgn}[\gamma - (1 - \gamma)(1 - 2s)]. \quad (12)$$

This confirms the result we have discussed before that, when  $\gamma > 1/2$ , the system is in the ordered phase  $S = +1$  for all values of  $m$  and  $s$  as each agent's satisfaction depends more on how it acts on other agents than on how the others act on it.

The case  $\gamma = 1/2$  is degenerate, giving

$$S = (1 - m) \text{sgn}(1 - s) + m \text{sgn } s. \quad (13)$$

As  $s \in [0, 1]$ , the above expression gives 1 for all values of  $m$  except for  $s = 0, 1$ , in which case linear relations with  $m$  are obtained. Because these are simply one dimensional lines on the borders of the diagram, they cannot be easily seen. This seems somewhat puzzling as completely selfish moral decisions from every single person are guaranteeing the well-being of the whole network even when both terms in the satisfaction have equal weights. The balance between the two competing terms  $U_i$  and  $W_i$  for  $\gamma = 1/2$  is however very

delicate. The strategy  $p_0 = 1$  guarantees that everyone has always the maximum value for its  $U_i$ , which means  $U_i/N \rightarrow 1$  in the thermodynamic limit. This implies that any small deviation from complete dissatisfaction concerning the way agents are being treated will tip the balance to the side of a satisfied population.

The most interesting cases are when  $\gamma < 1/2$ , which means that more importance is given to the way the agent is treated by others than to the way it treats other agents. Fig. 3 shows the result of the exact expressions and simulations for  $\gamma = 0.2$ . The simulations are run with a population size of  $N = 1000$  agents and averaged over 40 independently generated realizations of  $J$  and  $\pi$  configurations. The diagram at the center is the result of the simulation, the one at the left is the theoretical value for  $N = \infty$ . The difference between the two diagrams is due to finite size effects.

The diagram to the right represents the leading order theoretical correction for the system's finite size. Because  $\sigma^2$  in equation (A21) from appendix A is identically zero when  $p_0 = 0$ , we need to consider the variance coming from the average over the  $u_i$ 's. This leads to the expression

$$S(N) = (1 - m) \operatorname{erf} \left[ \frac{\gamma + (1 - \gamma)(1 - 2s)}{\sqrt{2[1 - (1 - 2s)^2]/N}} \right] + m \operatorname{erf} \left[ \frac{\gamma - (1 - \gamma)(1 - 2s)}{\sqrt{2[1 - (1 - 2s)^2]/N}} \right]. \quad (14)$$

One can see that a rich phase structure in the diagram develops with the appearance of three very distinctive bands. The central band represents the ordered phase  $S = 1$  and its width decreases with  $\gamma$ . This width can be calculated in the following way. Let us assume that  $\gamma = 1/2 - \epsilon$  where  $0 < \epsilon \leq 1/2$ . The ordered band requires the arguments of the two sign functions to be positive independently of the value of  $m$ , i.e.

$$\gamma + (1 - \gamma)(1 - 2s) > 0, \quad \gamma - (1 - \gamma)(1 - 2s) > 0. \quad (15)$$

This implies

$$\frac{2\epsilon}{1 + 2\epsilon} < s < \frac{1}{1 + 2\epsilon}, \quad (16)$$

and the bandwidth is therefore

$$\delta = \frac{1 - 2\epsilon}{1 + 2\epsilon}. \quad (17)$$

Notice that the area in which the whole network is fully satisfied with the distribution of moral decisions is larger than or *at most* equal to that in which it is not. This holds for *any* value of  $\gamma$  due to the symmetry of the lateral bands. This is a disheartening result as

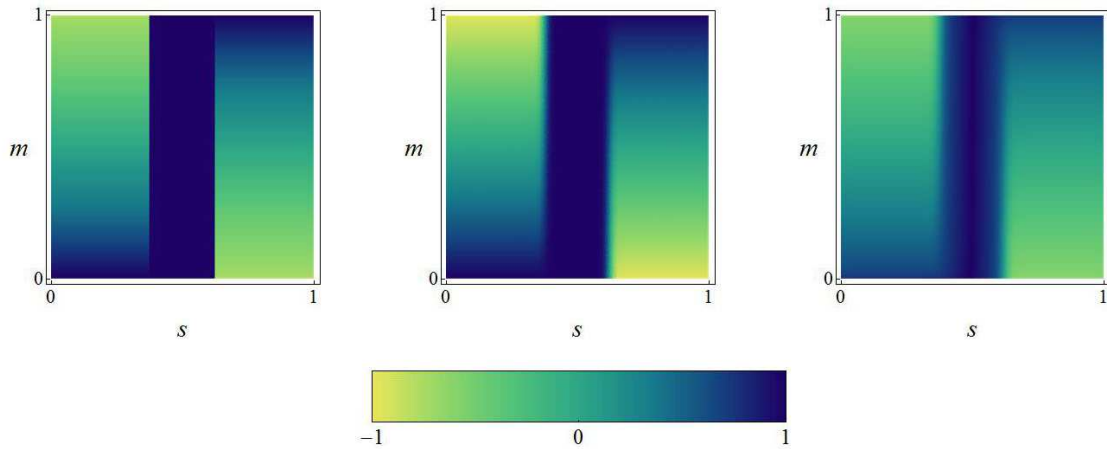


FIG. 3. Phase diagrams of selfish decisions for  $\gamma = 0.2$  (color online). Left: exact result for the infinite system; middle: computer simulation with  $N = 1000$  averaged over 40 independently generated configurations; right: leading order analytical approximation for  $N = 1000$ . The central band (online red) is totally ordered while the order parameter varies linearly from -1 to 1 in the lateral bands from top to bottom in the left band and from bottom to top in the right band.

it suggests that trying to derive moral decisions from rational principles that benefit society as a whole in an objective way might not be attainable. Such a principle would force one to accept that completely selfish decisions are morally correct, even when they are meant to do harm to those who do not want it. Of course this result is based on a simplified model, but it clearly illustrates that tying moral concepts to some rational measure of social well-being does not work in general.

The diagram is also interesting from the statistical physics point of view. By keeping  $m$  constant and varying  $s$  from zero to one, we pass through two discontinuous phase transitions. We start with a disordered phase which becomes ordered at the central band and then becomes disordered once again, but with an average satisfaction with the opposite sign. These transitions are shown in fig. 4. Inside each of the two disordered bands, by keeping  $s$  fixed and varying  $m$  in the interval  $[0, 1]$ , we have a continuous linear crossover between the two ordered phases with opposite signs of satisfaction.

Another interesting phase diagram is obtained by plotting the values of  $S$  in a  $\gamma \times s$  diagram, what is shown in fig. 5 for the fixed value  $m = 0.8$ . One can clearly see the quick growth of the central ordered band as the parameter  $\gamma$  increases from zero to  $1/2$ .

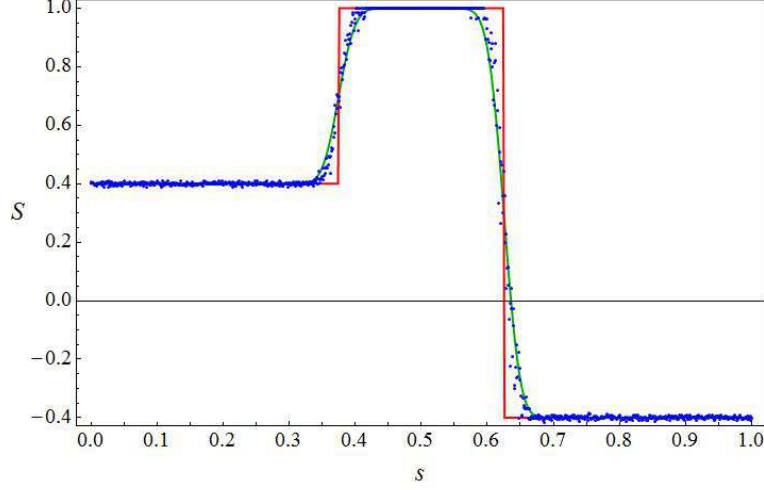


FIG. 4. The graph shows the order parameter  $S$  along a line of constant  $m = 0.3$  in the phase diagrams of fig. 3 (color online). The dots (online blue) represent the values of the simulations, the smooth line closer to them (online green) is the leading order finite size approximation and the non-smooth distribution (online red) is the infinite system. Notice that in this case  $\gamma = 0.2$ , giving a width of 0.25 for the ordered central band with edges at  $s = 0.375$  and  $s = 0.625$ .

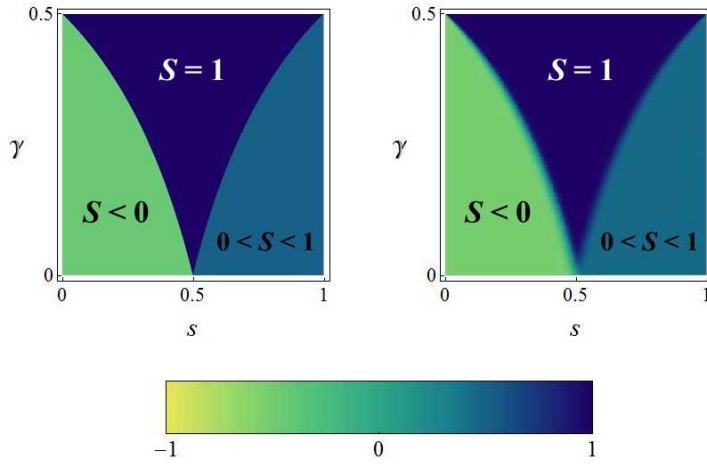


FIG. 5. Phase diagram (color online) showing the discontinuous transitions when  $\gamma$  and  $s$  are varied for constant  $m = 0.8$ . The left graph shows the result for an infinite system, while the right one shows the simulated diagram for  $N = 1000$ .

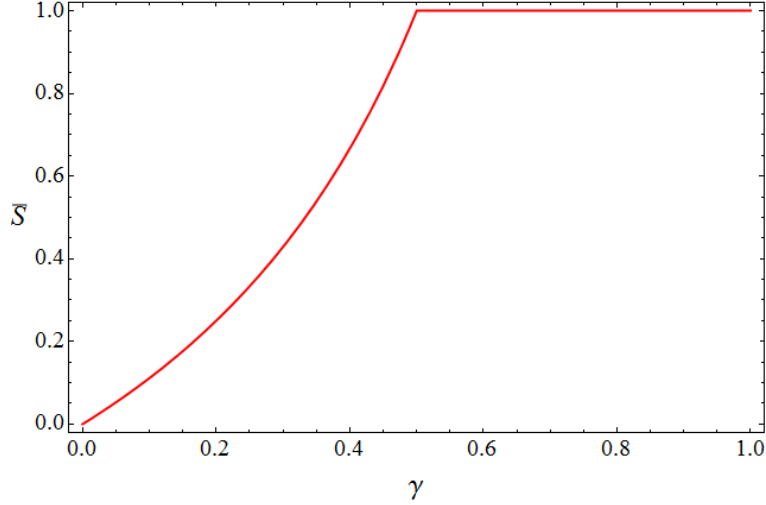


FIG. 6. Continuous phase transition from a disordered to an ordered phase at  $\gamma = 1/2$  with the order parameter taken as the average over the disorder in the personality vectors of the network satisfaction ( $\bar{S}$ ) for 0th order moral decisions.

Fig. 6 shows the continuous phase transition resulting from the change in the ordered band width at  $\gamma = 1/2$  by taking as the order parameter  $\bar{S}$ , the average of  $S$  over the disorder parameters  $s$  and  $m$ .

## B. 2nd Order Moral

From the moral point of view, one would be more inclined to accept that 2nd order moral decisions are better, or more considerate, choices than 0th order ones. The former is based on the idea of respecting others, while the latter bears no consideration for other's feelings. By construction, the model we are using has a symmetry connecting these two moral strategies. The substitution

$$\gamma \rightarrow 1 - \gamma, \quad u_i \rightarrow w_i, \quad J_{ij} \rightarrow J_{ji}, \quad (18)$$

keeps all  $\sigma_i$  the same and, therefore, also the average satisfaction  $S$ . The effect on  $S$  is equivalent if we change the last substitution by

$$p_0 = 1 \rightarrow p_2 = 1. \quad (19)$$

Therefore, if we change from selfish to empathetic moral decisions, the diagrams exchange their vertical and horizontal axes at the same time as  $\gamma \rightarrow 1 - \gamma$ . As in the selfish case, the

area of the diagram in which the satisfaction is positive is always larger than the negative one for all values of  $\gamma$ . If it was not for the fact that the same happens with the selfish behavior, that would be good news.

Although our model is quite simple, it is based on reasonable enough assumptions. Given the agreed concepts of morality, one can then see that two completely opposite moral decisions taken by the *whole* population guarantee the well-being of the network in the majority of the parameter space. This implies that, if one bases the concept of morality on “rational” arguments concerning the overall well-being, both behaviors should be considered morally right. None is more harmful to the society than the other.

This result seems paradoxical, but it arises from associating moral to satisfying the majority. Satisfying the majority is many times equivalent to ignoring or oppressing the minority, which is morally not acceptable. The paradox disappears if one recognizes that what can be associated with the definition of morally acceptable behavior is in fact the value of  $\gamma$ . We tend to consider behaviors to be morally acceptable when everyone is being respected by others, which is equivalent to count only the term  $W_i$  for the satisfaction. In other words, this is the situation when  $\gamma = 0$ . For the selfish strategy, a trivial calculation shows that this results in  $\bar{S} = 0$  while for the empathic strategy this gives  $\bar{S} = 1$ , the expected result for a morally accepted behavior.

There are many repressive moral beliefs in all societies which vary with time and that are not agreed by all individuals. For instance, the well-known repression of homosexual individuals in World War II in Britain, which was the probable cause of Alan Turing’s suicide, was considered a correct moral behavior at that time. In India and many other parts of the world, arranged marriages are still in practice even though the bride might not have a say in the final decision. Several religions have strict dressing codes and enforcing it, even violently in some cases, are considered as moral behavior even when it goes against the wishes of the individuals. However, these behaviors would all fall in the  $\bar{S} = 0$  region for  $\gamma = 0$ , which is a culturally independent criteria.

### C. 1st Order Moral

There is no symmetry connecting 0th and 2nd order moral decisions to 1st order ones. Taking only 1st order moral decisions creates a different phase diagram. The argument of

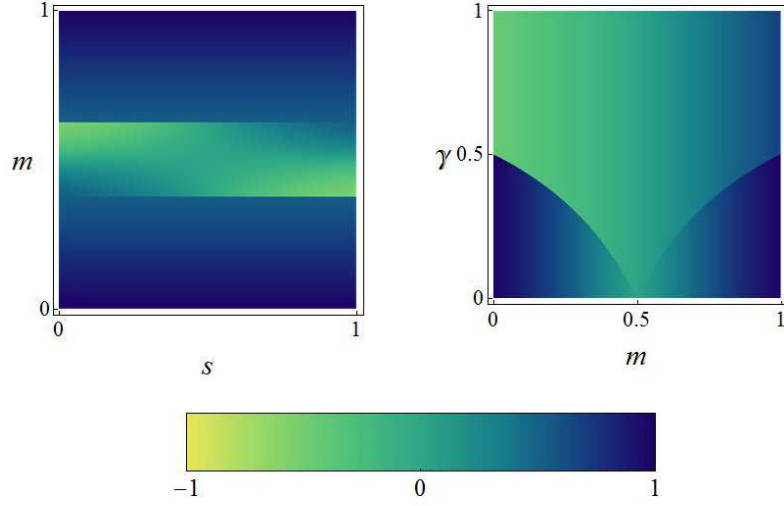


FIG. 7. Phase diagrams for the 1st order moral decisions (color online). Left:  $s \times m$  phase diagram for  $\gamma = 0.2$  showing the discontinuous transitions on the borders of the central band. Right:  $m \times \gamma$  diagram for  $s = 0.8$ . One can see the continuous transition in which the central band disappears completely at the value  $\gamma = 1/2$ .

the sign function simplifies to

$$\mu = \gamma uw + (1 - \gamma)(1 - 2m)w, \quad (20)$$

and the average satisfaction is then

$$S = (1 - 2m)\{(1 - s)\text{sgn}[\gamma + (1 - \gamma)(1 - 2m)] + s\text{sgn}[-\gamma + (1 - \gamma)(1 - 2m)]\}. \quad (21)$$

Analogous phase diagrams to the ones for the other strategies are given in fig. 7 which shows a very similar structure, but with different details of the phases and transitions. For instance, the  $s \times m$  diagram still presents a central band, which is however not totally ordered. It is also of opposite sign compared to the 0th and 2nd order strategies, with the central band indicating a disordered phase with negative satisfaction. The limits and size of the band are however the same as for the other strategies, which also implies the presence of a continuous phase transition in  $\bar{S}$  when  $\gamma$  is varied. This transition, depicted in fig. 8, is however between a partially ordered and a totally disordered phase, differently from those for the other strategies.



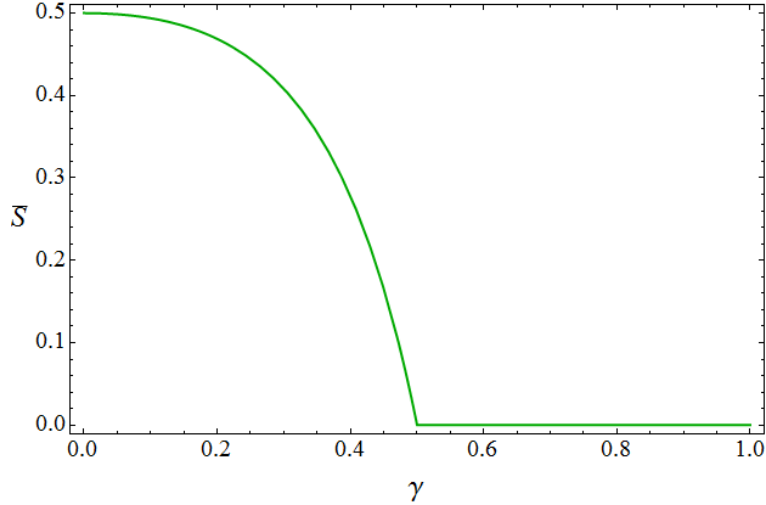


FIG. 8. Continuous phase transition from a partially ordered ( $\bar{S} = 1/2$ ) to a disordered phase at  $\gamma = 1/2$ .

This also shows that when  $\gamma = 0$  we have  $\bar{S} = 1/2$ . If we use the criteria suggested before, this is still a behavior which should not be considered morally acceptable, but would be *more* acceptable than the completely selfish decisions.

#### D. Mixed Strategy

For the sake of completeness, we will briefly consider a mixed strategy for which  $p_0 = p_1 = p_2 = 1/3$ . This strategy is unlikely to appear in real life as it has no well-defined rationale behind it. Fig. 9 shows a comparison between the  $n$ -order moral decisions and the mixed strategy. All strategies present continuous phase transitions in  $\bar{S}$  for  $\gamma = 1/2$  as is evident from the discontinuities in the derivatives at that point. Although the mixed strategy remains as the second best throughout all range of  $\gamma$ , it only attains  $\bar{S} = 1$  at the critical point.

As discussed before, it seems reasonable to associate some sort of *degree of morality* to each strategy by their values at  $\gamma = 0$ . It is then interesting to note that by randomly choosing each strategy with the same probability each time leads to a much higher moral degree than the 0th and 1st order strategies. One might be tempted to argue that some respect once in a while is much better than none. This striking difference can be better understood by looking at the phase diagrams  $s \times m$  for all strategies at  $\gamma = 0$  given in fig.

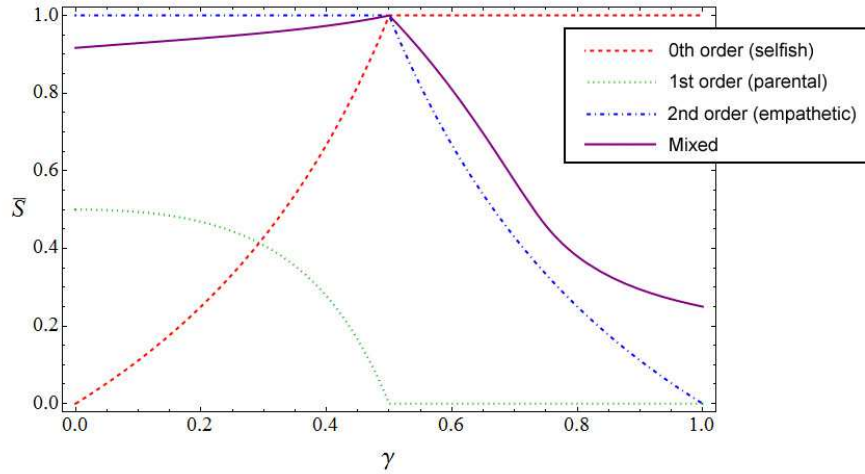


FIG. 9. Continuous phase transitions on  $\bar{S} = 1/2$  for the four strategies at  $\gamma = 1/2$ .

10.

#### IV. CONCLUSIONS

In this work we analyzed a model of agents interacting in a fully connected network by taking moral decisions. The agent's decisions were constrained by an increased level of empathy based on the stages of Piaget's theory of cognitive development. Within this model, we defined a measure of emotional satisfaction related to the fulfillment of the agents' personal desires to characterize the phases of the network as the parameters of the model are varied. These desires were modeled by local random binary fields at each site of the network and were called *personality vectors* as they represent personality traits of the agents.

The introduced model is exactly solvable. We were able to calculate analytically the value of the average satisfaction of the network in the limit of an infinite number of agents and also its leading order contribution in powers of  $1/N$ . Using the average satisfaction as an order parameter, we then were able to find the model's phase diagrams and to identify the existence of both continuous and discontinuous transitions triggered by the variation of the disorder parameters.

We chose to analyze four different specific strategies of moral decisions in this work. In the first three, agents take decisions according to each stage in Piaget's theory, a three-steps hierarchy to which we gave the name of *Piaget's Ladder*. Each step in this ladder

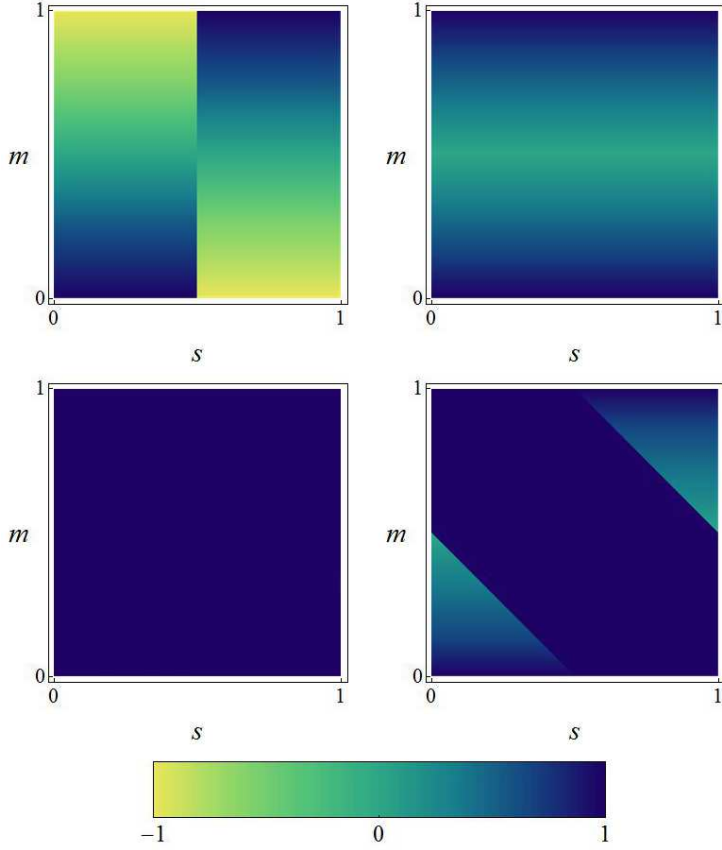


FIG. 10. Phase diagrams for the four strategies at  $\gamma = 0$  (color online). From left to right. Top row: 0th order, 1st order. Bottom row: 2nd order, mixed.

is associated with an  $n$ -th order moral. The 0th order corresponds to completely selfish decisions and the 2nd order to completely empathetic ones. The 1st order corresponds to an intermediate case. In addition to these three strategies, we also analyzed the case in which each agent takes a decision by randomly choosing one of those strategies at each time.

Interestingly, there is a symmetry between the 0th and 2nd order moral decisions resulting in the fact that the selfish strategy guarantees the *average* well-being of the network as well as the empathetic one. Although the model is very simple, its assumptions are reasonable enough to indicate that moral beliefs cannot be simply based on rational minimization of some energy function as two strategies that would clearly be considered morally opposite by most people lead to the same phase diagrams. One possibility is that the parameters of the disorder are tuned in humans to values for which a species-wide agreement on moral

concepts can be derived. One solution to this conundrum suggested by us was that a moral degree might be associated not to the overall satisfaction, but to its particular case in which the satisfaction of others is more important than ours. This leads to a sensible classification when all four strategies are compared. Of course, a more realistic view, as that of MFT, would require a more sophisticated set of parameters, but this is out of the scope of the present analysis.

The 0th and 2nd order strategies minimize the Hamiltonian in the appropriate domains of the parameter  $\gamma$  which measures the relative importance of the two terms in the satisfaction. As expected, the 1st order and the mixed strategies underperform those strategies in their optimal ranges, although the mixed strategy seems to be the one which remains more acceptable in a wider range of  $\gamma$ .

From the technical point of view, the present model has many interesting properties. It is simple enough to be easily interpretable and completely solvable. At the same time, it presents a range of very interesting phase diagrams and transitions. We have discussed only some basic features of these transitions due to size constraints of this paper, choosing to focus on the interpretation of these results. In a forthcoming paper, we intend to explore the interesting mathematical structure of this model in more detail.

Finally, there are many directions in which this model can be extended in order to include more realistic behavior. For instance, different network topologies can be used. One could also devise a scenario in which agents try to infer by a learning algorithm the desires of others. This would bring dynamics into the model and require a more sophisticated treatment which we will leave for a future study. The personality vector also can be extended either to a higher dimensionality to include more realistic personality variations or to continuous instead of binary values. We intend to explore these extensions in future works.

## ACKNOWLEDGMENTS

I would like to thank Dr Juan Neirotti for very useful discussions and comments.

- 
- [1] P. E. Tetlock and D. Gardner, *Superforecasting: The art and science of prediction* (Signal, 2015).

- [2] C. Castellano, S. Fortunato, and V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).
- [3] D. M. Abrams, H. A. Yapel, and R. J. Wiener, *Phys. Rev. Lett.* **107**, 088701 (2011).
- [4] R. Vicente, A. Susemihl, J. P. Jericó, and N. Caticha, *Physica A: Statistical Mechanics and its Applications* **400**, 124 (2014).
- [5] G. J. Ross and T. Jones, *Phys. Rev. E* **91**, 062809 (2015).
- [6] S. Galam, *Sociophysics: a physicist's modeling of psycho-political phenomena* (Springer Science & Business Media, 2012).
- [7] J. Reichardt, R. Alamino, and D. Saad, *PloS one* **6**, e21282 (2011).
- [8] L. A. Martinez-Vaquero and J. A. Cuesta, *Phys. Rev. E* **90**, 022805 (2014).
- [9] N. C. Clementi, J. A. Revelli, and G. J. Sibona, *Phys. Rev. E* **92**, 012816 (2015).
- [10] H. S. Sugiarto, N. N. Chung, C. H. Lai, and L. Y. Chew, *Phys. Rev. E* **91**, 062804 (2015).
- [11] R. V. Solé, *Phase Transitions* (Princeton University Press, 2011).
- [12] D. S. Wilson, *Darwin's Cathedral: Evolution, Religion, and the Nature of Society* (University Of Chicago Press, Chicago, 2003).
- [13] S. Okasha, *Evolution and the Levels of Selection* (Oxford University Press, Oxford, 2006).
- [14] R. H. Schonmann, R. Vicente, and N. Caticha, arXiv:1208.0863v2 [q-bio.PE] (2012).
- [15] J. Haidt, *Science* **316**, 998 (2007).
- [16] A. H. Maslow, *Psychological Review* **50**, 370 (1943).
- [17] J. Piaget, *The Origin of Intelligence in Children* (International Universities Press, New York, 1965).
- [18] L. Oakley, *Cognitive development* (Routledge, Hove, UK, 2004).
- [19] D. Ariely, *Predictably Irrational* (HarperCollins, New York, USA, 2008).
- [20] N. L. Quenk, *Essentials of Myers-Briggs type indicator assessment*, Vol. 66 (John Wiley & Sons, 2009).

## Appendix A: Analytical Calculation of the Average Satisfaction

We can explicitly write the expression for the average satisfaction  $S$  as

$$S = \langle \sigma_k \rangle_{\mathbf{u}, \mathbf{w}, J} = \sum_{\{u_i\}} \sum_{\{w_i\}} \sum_{\{J_{ij}\}} \left[ \prod_i \mathcal{P}(u_i) \mathcal{P}(w_i) \right] \left[ \prod_{\substack{i,j \\ i \neq j}} \mathcal{P}(J_{ij} | u_i, w_i, u_j, w_j) \right] \times \text{sgn} \left[ \frac{1}{N} \left( \gamma u_k \sum_{l \neq k} J_{kl} + (1 - \gamma) w_k \sum_{l \neq k} J_{lk} \right) \right]. \quad (\text{A1})$$

We start by doing the average over  $J$ . The argument of the sign contains averages over  $N$  variables  $J_{ij}$ . Although they are not identically distributed, the fact that they are independent given  $\mathbf{u}$  and  $\mathbf{w}$  allows us to use an extension of the Central Limit Theorem (CLT). In the following, we explicitly present this extension.

By means of a Dirac delta distribution, one can write the above equation as

$$S = \left\langle \int \frac{dx d\hat{x}}{2\pi} e^{ix\hat{x}} (\text{sgn } x) \Gamma(\hat{x}, \mathbf{u}, \mathbf{w}, \gamma) \right\rangle_{\mathbf{u}, \mathbf{w}}, \quad (\text{A2})$$

with

$$\Gamma = \sum_{\{J_{ij}\}} \left[ \prod_{\substack{i,j \\ i \neq j}} \mathcal{P}(J_{ij} | u_i, w_i, u_j, w_j) \right] \exp \left\{ -\frac{i\hat{x}}{N} \left[ \gamma u_k \sum_{l \neq k} J_{kl} + (1 - \gamma) w_k \sum_{l \neq k} J_{lk} \right] \right\}. \quad (\text{A3})$$

The average can now be factorized and written as

$$\prod_{l \neq k} \Lambda_{lk}^1 \Lambda_{lk}^2 = \exp \left\{ \sum_{l \neq k} (\ln \Lambda_{lk}^1 + \ln \Lambda_{lk}^2) \right\}, \quad (\text{A4})$$

where

$$\Lambda_{lk}^1 \equiv \sum_{J_{kl}} \mathcal{P}(J_{kl}) \exp \left\{ -\frac{i\hat{x}}{N} \gamma u_k J_{kl} \right\}, \quad (\text{A5})$$

$$\Lambda_{lk}^2 \equiv \sum_{J_{lk}} \mathcal{P}(J_{lk}) \exp \left\{ -\frac{i\hat{x}}{N} (1 - \gamma) w_k J_{lk} \right\}, \quad (\text{A6})$$

with  $k$  a fixed index.

Given that  $J$  is a binary matrix, we can rewrite the probability distributions as

$$\mathcal{P}(J_{ij} | \pi_i, \pi_j) = p_0 \left( \frac{1 + J_{ij} u_i}{2} \right) + p_1 \left( \frac{1 + J_{ij} w_i}{2} \right) + p_2 \left( \frac{1 + J_{ij} w_j}{2} \right), \quad (\text{A7})$$

which gives

$$\Lambda_{lk}^1 = \cos \left[ \frac{\gamma \hat{x}}{N} \right] - i(p_0 + u_k w_k p_1 + u_k w_l p_2) \sin \left[ \frac{\gamma \hat{x}}{N} \right] \quad (\text{A8})$$

$$\Lambda_{lk}^2 = \cos \left[ \frac{(1-\gamma) \hat{x}}{N} \right] - i(u_l w_k p_0 + w_l w_k p_1 + p_2) \sin \left[ \frac{(1-\gamma) \hat{x}}{N} \right]. \quad (\text{A9})$$

Expanding the cosines, sines and logarithms up to order  $1/N^2$ , we get

$$\ln \Lambda_{lk}^1 \approx -\frac{\gamma^2 \hat{x}^2}{2N^2} \left[ 1 - (\lambda_{kl}^1)^2 \right] - i \frac{\gamma \hat{x}}{N} \lambda_{kl}^1, \quad (\text{A10})$$

$$\ln \Lambda_{lk}^2 \approx -\frac{(1-\gamma)^2 \hat{x}^2}{2N^2} \left[ 1 - (\lambda_{kl}^2)^2 \right] - i \frac{(1-\gamma) \hat{x}}{N} \lambda_{kl}^2, \quad (\text{A11})$$

where

$$\lambda_{kl}^1 = p_0 + u_k w_k p_1 + u_k w_l p_2, \quad (\text{A12})$$

$$\lambda_{kl}^2 = u_l w_k p_0 + w_l w_k p_1 + p_2. \quad (\text{A13})$$

By defining

$$\mu \equiv \frac{1}{N} \sum_{l \neq k} [\gamma \lambda_{kl}^1 + (1-\gamma) \lambda_{kl}^2], \quad (\text{A14})$$

$$\sigma^2 \equiv \frac{1}{N^2} \sum_{l \neq k} \left\{ \gamma^2 \left[ 1 - (\lambda_{kl}^1)^2 \right] + (1-\gamma)^2 \left[ 1 - (\lambda_{kl}^2)^2 \right] \right\}, \quad (\text{A15})$$

we can write

$$\begin{aligned} S &= \left\langle \int \frac{dx d\hat{x}}{2\pi} e^{-\frac{\hat{x}^2 \sigma^2}{2} + i\hat{x}(x-\mu)} (\text{sgn } x) \right\rangle_{\mathbf{u}, \mathbf{w}} \\ &= \left\langle \text{erf} \left( \frac{\mu}{\sqrt{2\sigma^2}} \right) \right\rangle_{\mathbf{u}, \mathbf{w}}. \end{aligned} \quad (\text{A16})$$

Notice that this is the extension of the CLT that we alluded to. The average over  $J$  became an average over a Gaussian distributed variable  $x$  with mean  $\mu$  and variance  $\sigma^2$  which are what we would obtain by calculating the mean and variance of each  $J_{ij}$  and doing the appropriate linear combination or, using the notation of equation (2),

$$\mu = \frac{1}{N} \sum_{l \neq k} [\gamma u_k \langle J_{kl} \rangle + (1-\gamma) w_k \langle J_{lk} \rangle], \quad (\text{A17})$$

$$\sigma^2 = \frac{1}{N^2} \sum_{l \neq k} [\gamma^2 (1 - \langle J_{kl} \rangle^2) + (1-\gamma)^2 (1 - \langle J_{lk} \rangle^2)], \quad (\text{A18})$$

where

$$\langle J_{ij} \rangle = p_0 u_i + p_1 w_i + p_2 w_j. \quad (\text{A19})$$

The expressions for  $\mu$  and  $\sigma^2$  can be further simplified for  $N \rightarrow \infty$

$$\begin{aligned}\mu &= \frac{1}{N} \sum_{l \neq k} [\gamma(p_0 + u_k w_k p_1 + u_k w_l p_2) + (1 - \gamma)(u_l w_k p_0 + w_l w_k p_1 + p_2)] \\ &= p_0[\gamma + (1 - \gamma)\bar{u}w_k] + p_1[\gamma u_k w_k + (1 - \gamma)\bar{w}w_k] + p_2[\gamma \bar{w}u_k + (1 - \gamma)],\end{aligned}\tag{A20}$$

and

$$\begin{aligned}\sigma^2 &= \frac{1}{N^2} \sum_{l \neq k} \{ \gamma^2 [1 - (p_0 + u_k w_k p_1 + u_k w_l p_2)^2] + (1 - \gamma)^2 [1 - (u_l w_k p_0 + w_l w_k p_1 + p_2)^2] \} \\ &= \frac{1}{N} \{ [\gamma^2 + (1 - \gamma)^2] (1 - p_0^2 - p_1^2 - p_2^2) - 2p_0 p_1 [\gamma^2 u_k w_k + (1 - \gamma)^2 \bar{C}] \\ &\quad - 2p_0 p_2 [\gamma^2 u_k \bar{w} + (1 - \gamma)^2 \bar{u} w_k] - 2p_1 p_2 [\gamma^2 + (1 - \gamma)^2] w_k \bar{w} \}.\end{aligned}\tag{A21}$$

where we introduced the definitions

$$\bar{u} = \frac{1}{N} \sum_{l \neq k} u_l, \quad \bar{w} = \frac{1}{N} \sum_{l \neq k} w_l, \quad \bar{C} = \frac{1}{N} \sum_{l \neq k} u_l w_l.\tag{A22}$$

The CLT can now be directly applied in its original form to the above variables, resulting in

$$S = \left\langle \operatorname{erf} \left( \frac{\mu}{\sqrt{2\sigma^2}} \right) \right\rangle_{u_k, w_k, \bar{u}, \bar{w}, \bar{C}},\tag{A23}$$

where the hatted variables are distributed according to the following Gaussian distributions

$$\bar{u} \sim \mathcal{N}(\bar{u} | \langle u_i \rangle, \sigma_u^2),\tag{A24}$$

$$\bar{w} \sim \mathcal{N}(\bar{w} | \langle w_i \rangle, \sigma_w^2),\tag{A25}$$

$$\bar{C} \sim \mathcal{N}(\bar{C} | \langle u_i \rangle \langle w_i \rangle, \sigma_{uw}^2),\tag{A26}$$

where

$$\mathcal{N}(y | \mu_y, \sigma_y^2) \equiv \frac{e^{-\frac{(y - \mu_y)^2}{2\sigma_y^2}}}{\sqrt{2\pi\sigma_y^2}},\tag{A27}$$

and

$$\langle u_i \rangle = (1 - 2s), \quad \sigma_u^2 = 1 - \langle u_i \rangle^2 = 4s(1 - s),\tag{A28}$$

$$\langle w_i \rangle = (1 - 2m), \quad \sigma_w^2 = 1 - \langle w_i \rangle^2 = 4m(1 - m),\tag{A29}$$

$$\sigma_{u_i w_i}^2 = 1 - \langle u_i \rangle^2 \langle w_i \rangle^2.\tag{A30}$$



We do not need to carry the index  $k$  anymore and therefore we write  $u$  and  $w$  instead of  $u_k$  and  $w_k$ . In the limit  $N \rightarrow \infty$ , the above Gaussians become Dirac deltas in their means and the error function becomes the sign of its argument, which gives our final expression

$$S = \langle \text{sgn } \mu \rangle_{u,w}, \quad (\text{A31})$$

with

$$\mu = p_0[\gamma + (1 - \gamma)(1 - 2s)w] + p_1[\gamma uw + (1 - \gamma)(1 - 2m)w] + p_2[\gamma(1 - 2m)u + (1 - \gamma)]. \quad (\text{A32})$$